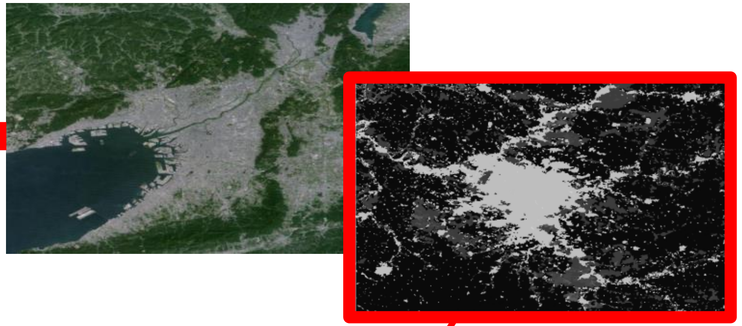
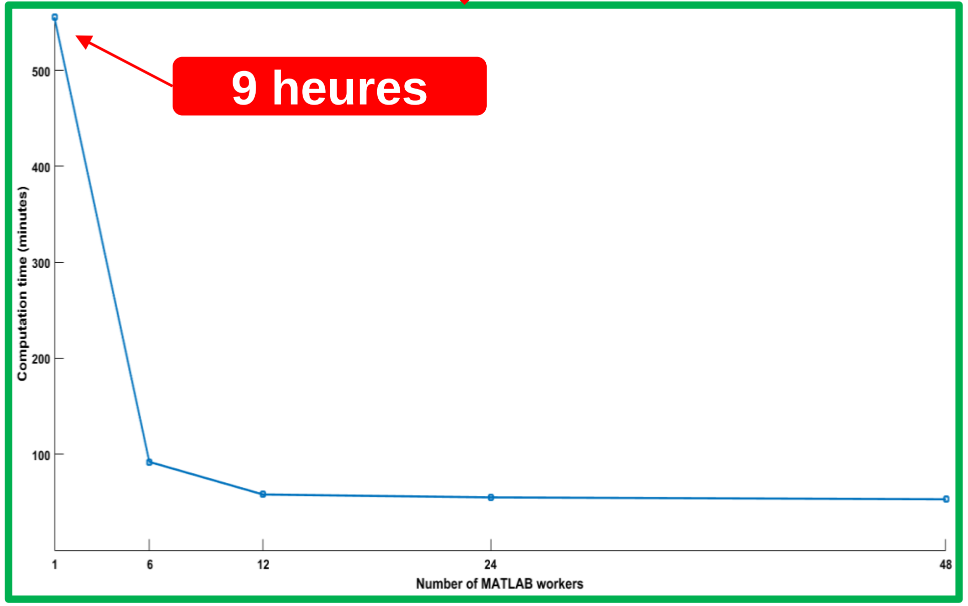
Parallel Computing & Big Data with MATLAB

Lyamine Hedjazi Application Engineer
Nadia Bedjaoui Academic Engineer

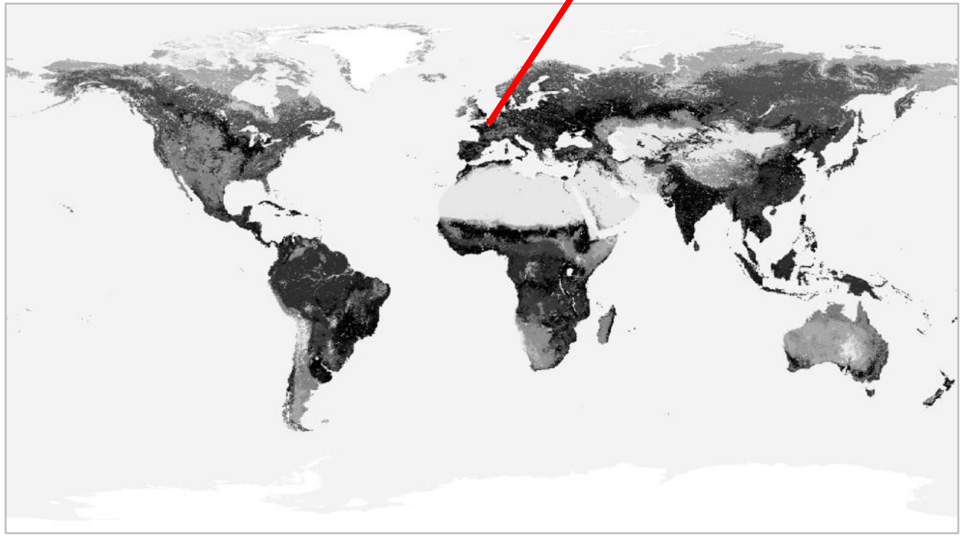
Drastic Time Analysis Reduction of Urban Areas

Parallel Image Processing Approach





2 minutes



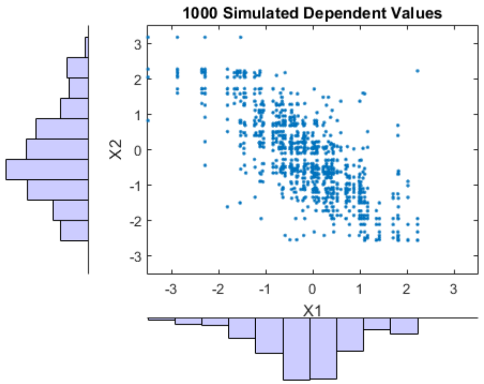
Parallel Computing can help you to…

Increase Productivity

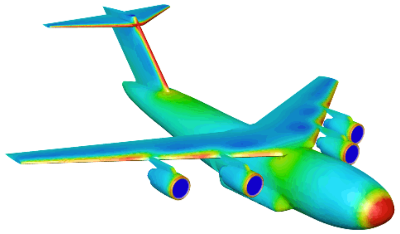
- **Reduce computation times**
  - Accelerate your simulations
  - By taking advantage of unused CPU cores and/or GPUs

Increase Accuracy

- **Run more simulations** within the same timeframe
  - Test more values for parameter calibration and design optimization
  - 2x more runs in Monte Carlo simulations ⇨ 41% increase in accuracy



- **Run larger simulations** within the same timeframe
  - Use finer meshes for Finite Element Analysis
  - Not enough memory on a single computer? ⇨ Distribute the mesh on a cluster!



How much faster can my computations run?

- For an application that is well-suited for parallel computing…
  - Monte Carlo simulations
  - Parameter calibration / Design optimization
  - Linear algebra on very large matrices
  - Processing of large images and/or large collection of images

- … the speed-up mainly depends on the **number of available CPU cores**

Laptop
1.5x to 3x

Workstation
1.5x to 20x

HPC cluster (multiple computers)
10x to 1000x or even more!

MathWorks Solutions for Parallel Computing

Parallel Computing Toolbox

- MATLAB Toolbox
- Allows to write parallel applications by giving access to parallel programming constructs
- Take advantage of **local** CPU cores and GPUs

Laptop
1.5x to 3x

Workstation
1.5x to 20x

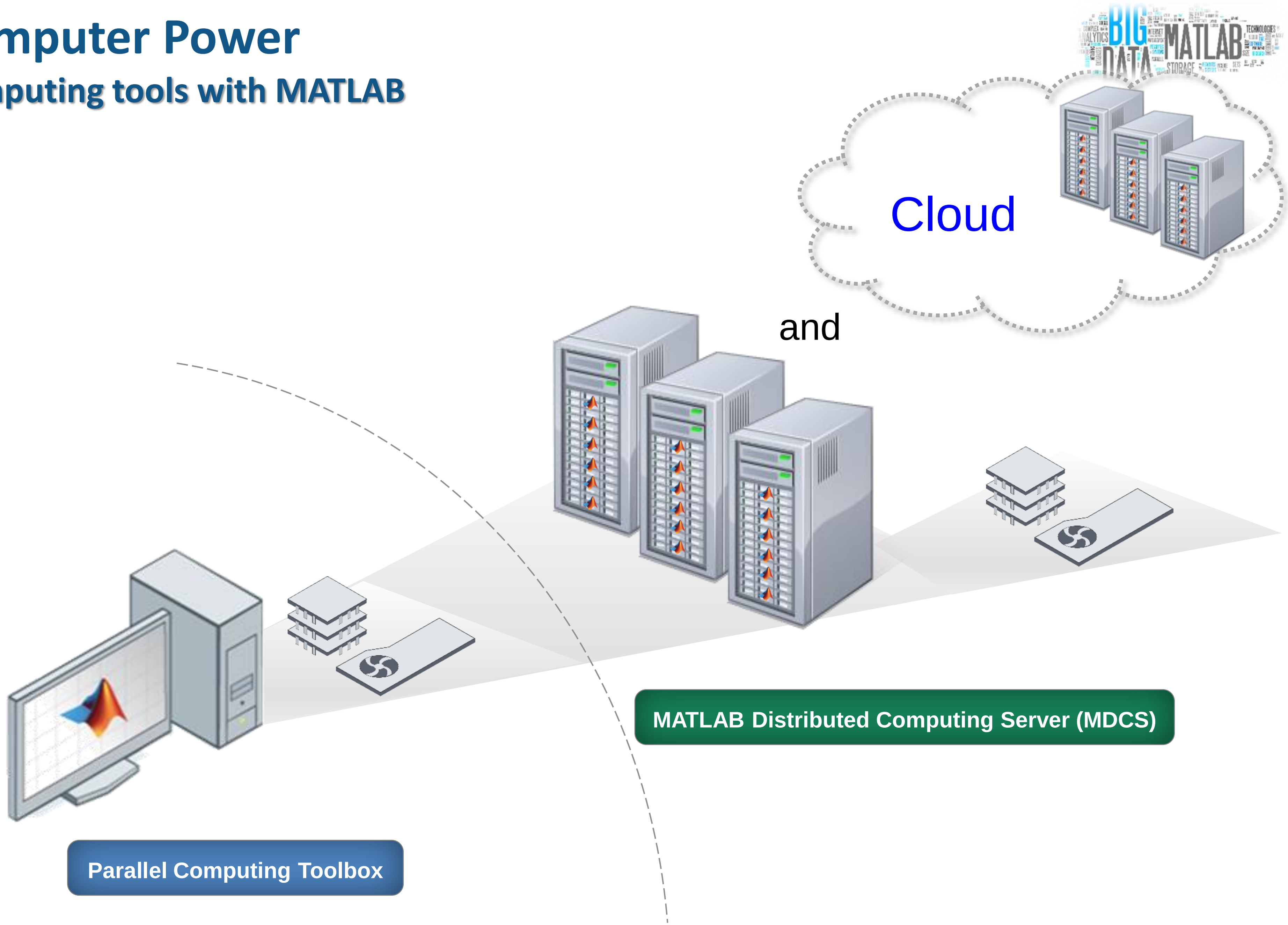
MATLAB Distributed Computing Server

- Standalone product
- Run computations on remote resources
- With MDCS, a single application can be executed on multiple computers

HPC cluster (multiple computers)
10x to 1000x or even more!

Scale Computer Power

Parallel Computing tools with MATLAB



Scaling-up locally

Use the full processing power of desktops (CPUs, GPUs)

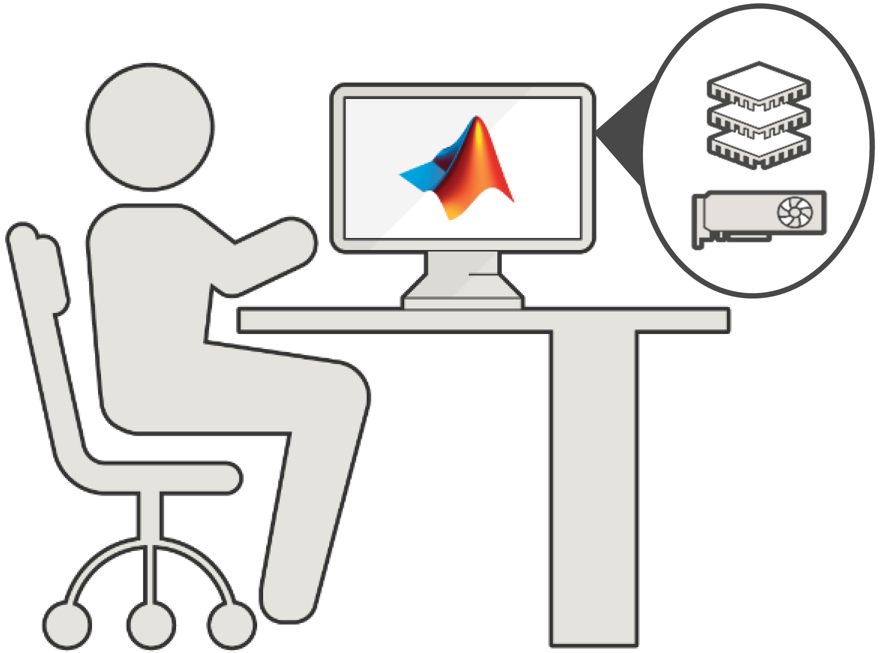
- Leverage GPU accelerator

```
d=gpuDevice()
```

- Leverage Multicore CPU

```
p=parpool('local');
```

Parallel Computing Toolbox



Programming Parallel Applications

Leverage GPU Accelerators

Ease of Use

- Built-in GPU support
  - Neural Network Toolbox
  - Many signal & image processing features
- GPU-accelerated functions
  - Transfer data to the GPU and run computations on the GPU
  - **Overloaded** functions, no need to rewrite your code
  - Most linear algebra functions, FFT, statistics…
- Import CUDA code into MATLAB
  - If you have written C/C++/FORTRAN code that uses the CUDA library, you can import it in MATLAB using **MEX files**
- Important remarks
  - Only NVIDIA GPUs are supported
  - Make sure to use GPUs with full double precision support (NVIDIA Tesla)

Greater Control

Leveraging your parallel computing resources

Multicore CPUs and systems

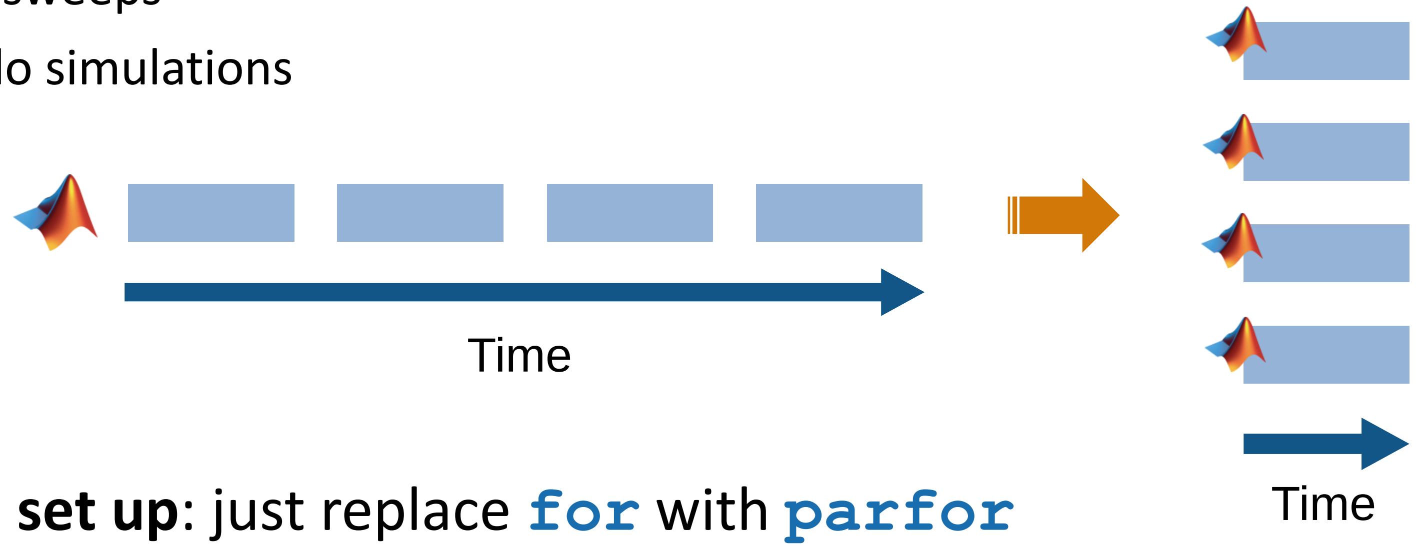
Ease of Use

- Built-in Parallel Computing support
  - Many toolboxes can benefit from parallel speed-up
    - Optimization & Global Optimization Toolboxes
    - Statistics and Machine Learning Toolbox
    - Neural Network Toolbox
    - Image Processing Toolbox & Computer Vision System Toolbox
    - …
  - Check the documentation for a **UseParallel** property
  - Or refer to www.mathworks.com/builtin-parallel-support
- **parfor** (parallel-for loops)
  - Speed up for loops with **independent** iterations
  - Monte Carlo simulations, parameter calibration

Greater Control

Parallel-for Loops

- Parallel-for loops used to **speed-up for loops** with **independent** iterations
  - Parameter sweeps
  - Monte Carlo simulations



- **Very easy to set up**: just replace `for` with `parfor`

- Ideal use cases for parallel computing
  - Low overhead, no communication between tasks
  - Excellent scalability and speed-ups

Leveraging your parallel computing resources

Multicore CPUs and systems

Ease of Use

- Built-in Parallel Computing support
  - Many toolboxes can benefit from parallel speed-up
    - Optimization & Global Optimization Toolboxes
    - Statistics and Machine Learning Toolbox
    - Neural Network Toolbox
    - Image Processing Toolbox & Computer Vision System Toolbox
    - …
  - Check the documentation for a **UseParallel** property
  - Or refer to www.mathworks.com/builtin-parallel-support
- **parfor** (parallel-for loops)
  - Speed up for loops with **independent** iterations
  - Monte Carlo simulations, parameter calibration
- MPI-like programming constructs
  - **spmd** blocks
  - **labSend**, **labReceive**, **gop**

Greater Control

spmd blocks

```
spmd
    % single program across workers
end
```

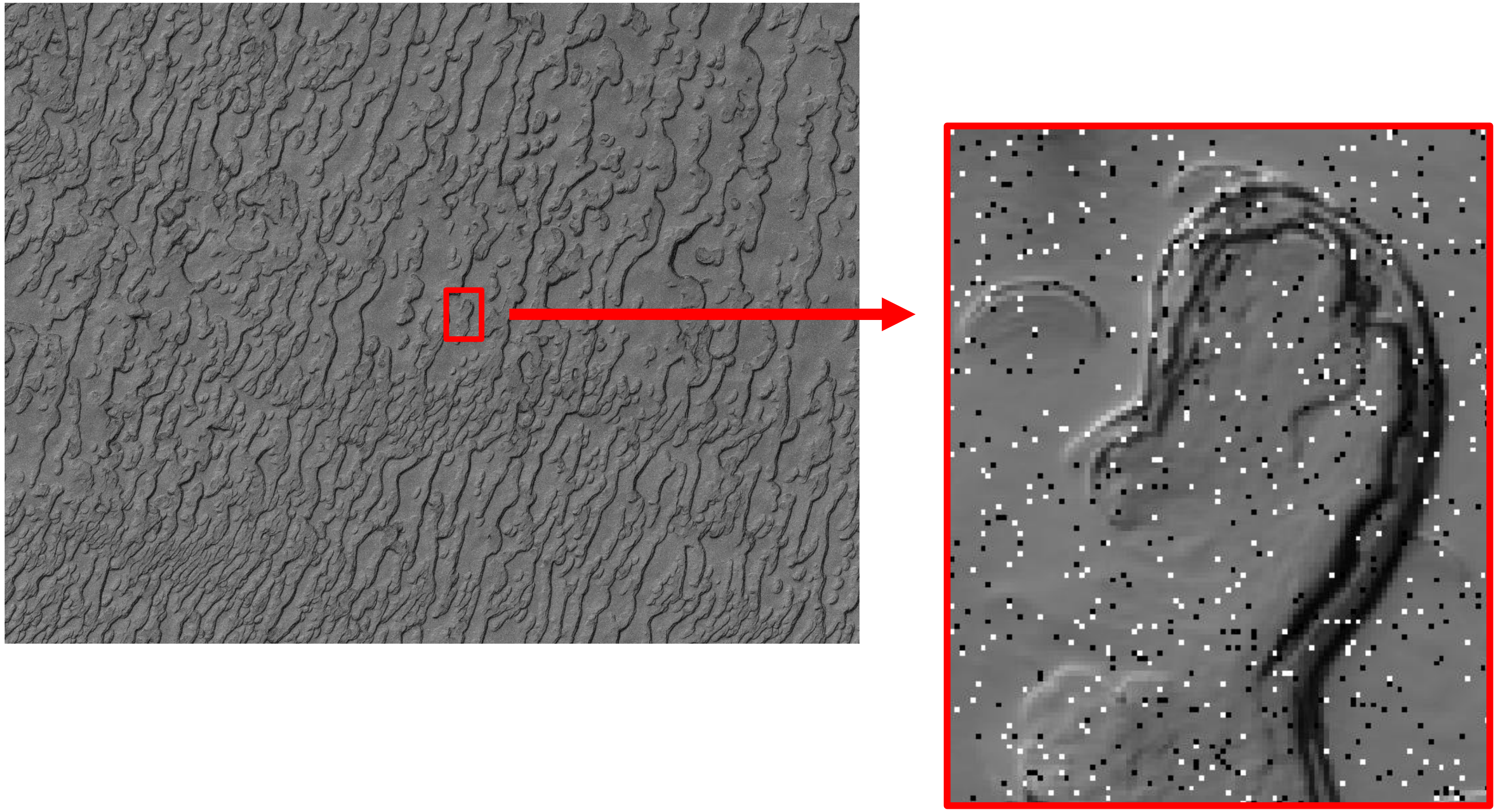
- Mix parallel and serial code in the same function
- Run on a pool of MATLAB resources
- **S**ingle **P**rogram runs simultaneously across workers
- **M**ultiple **D**ata spread across multiple workers

```
spmd
    a = labindex + numlabs;
    disp(['a = ',
num2str(a)]);
end
```

```
>> Lab 1
            a = 5
   Lab 2
            a = 6
   Lab 3
            a = 7
   Lab 4
            a = 8
```

Communicating tasks – Denoising a Hi-Res Image

- **Goal**: denoising a large image using median filtering



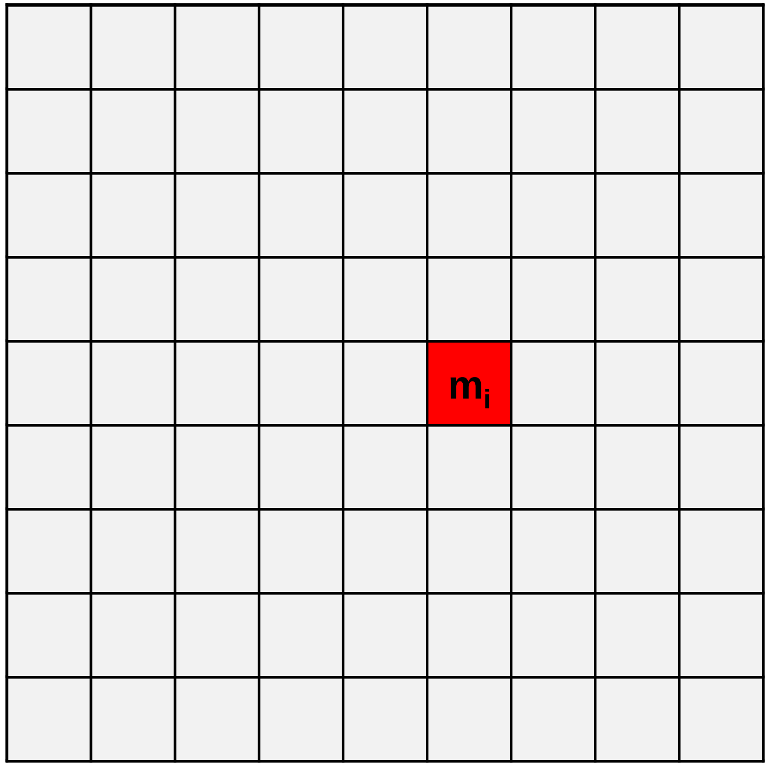
Communicating tasks – Denoising a Hi-Res Image

Median filtering

- Useful tool for removing ‘salt and pepper’ noise
- Filter is applied pixel by pixel

- **Algorithm**
  - for each pixel $\mathbf{p_i}$ in the image
  - take a square subset of width **w** around $\mathbf{p_i}$
  - compute the median value $\mathbf{m_i}$ of this subset
  - replace the value of $\mathbf{p_i}$ with $\mathbf{m_i}$

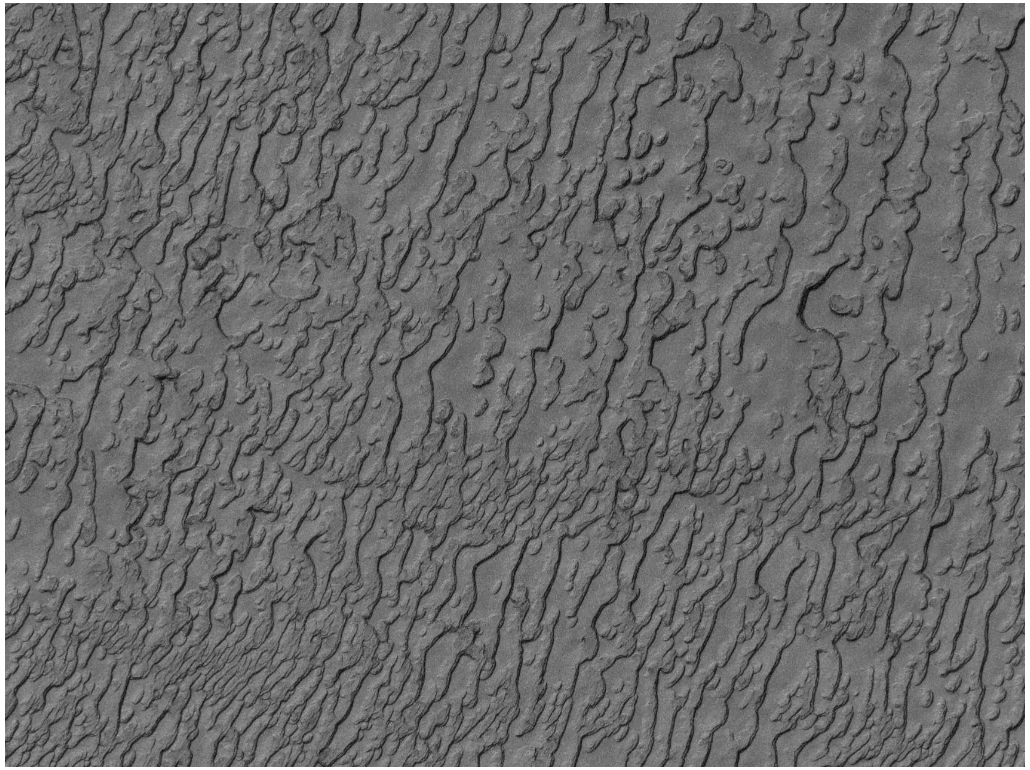
- **MATLAB functions**
  - `medfilt1` for the 1-D case
  - `medfilt2` for the 2-D case



$\mathbf{m_i}$ = median(yellow square)

Communicating tasks – Denoising a Hi-Res Image

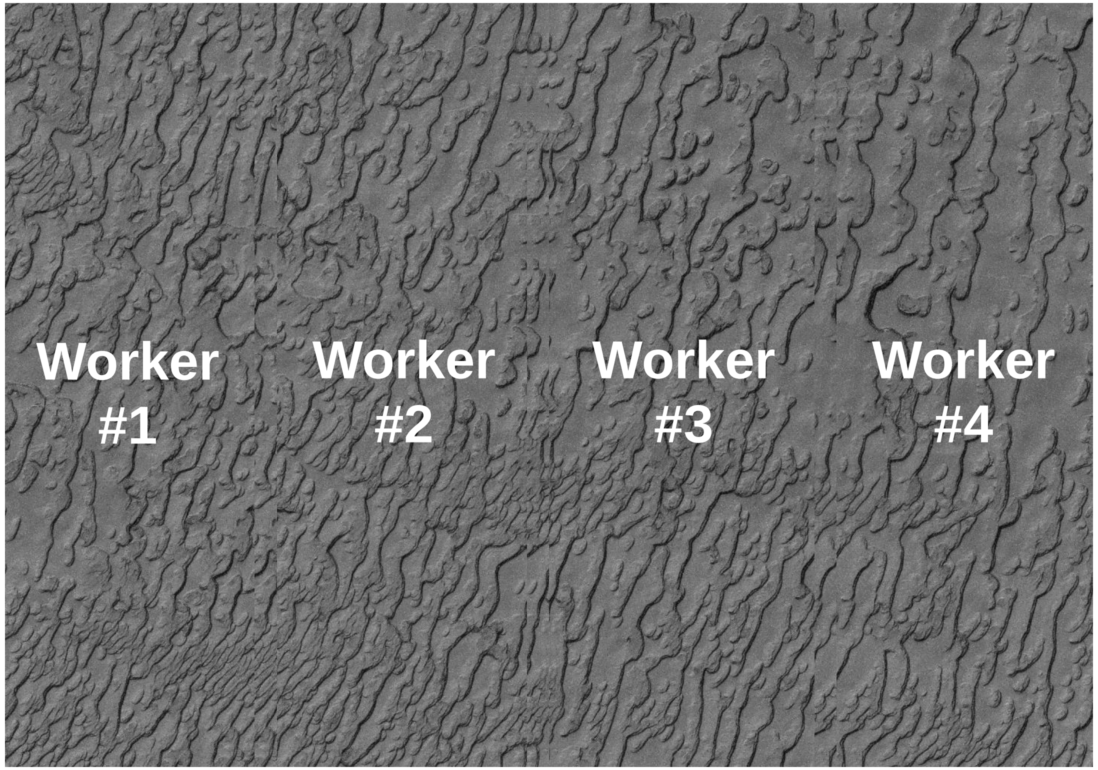
Parallelization technique on multicore CPU(s)



Communicating tasks – Denoising a Hi-Res Image

Parallelization technique on multicore CPU(s)

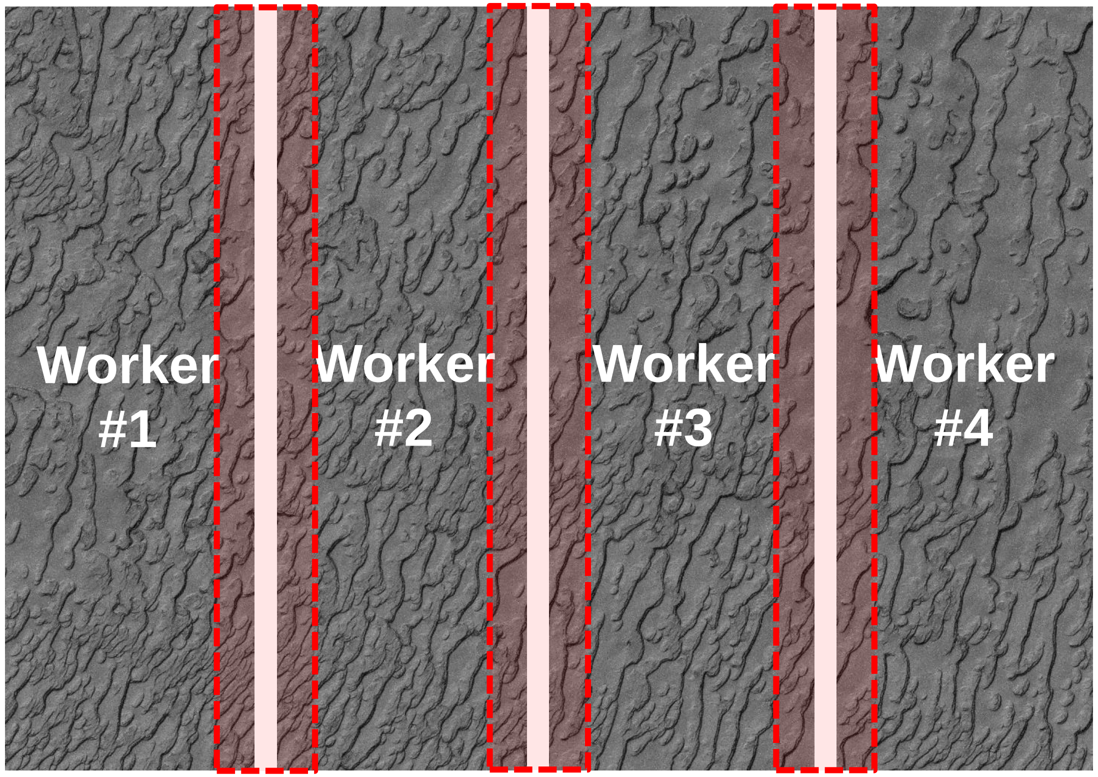
- Divide the image in multiple stripes, each processed by a different worker



Communicating tasks – Denoising a Hi-Res Image

Parallelization technique on multicore CPU(s)

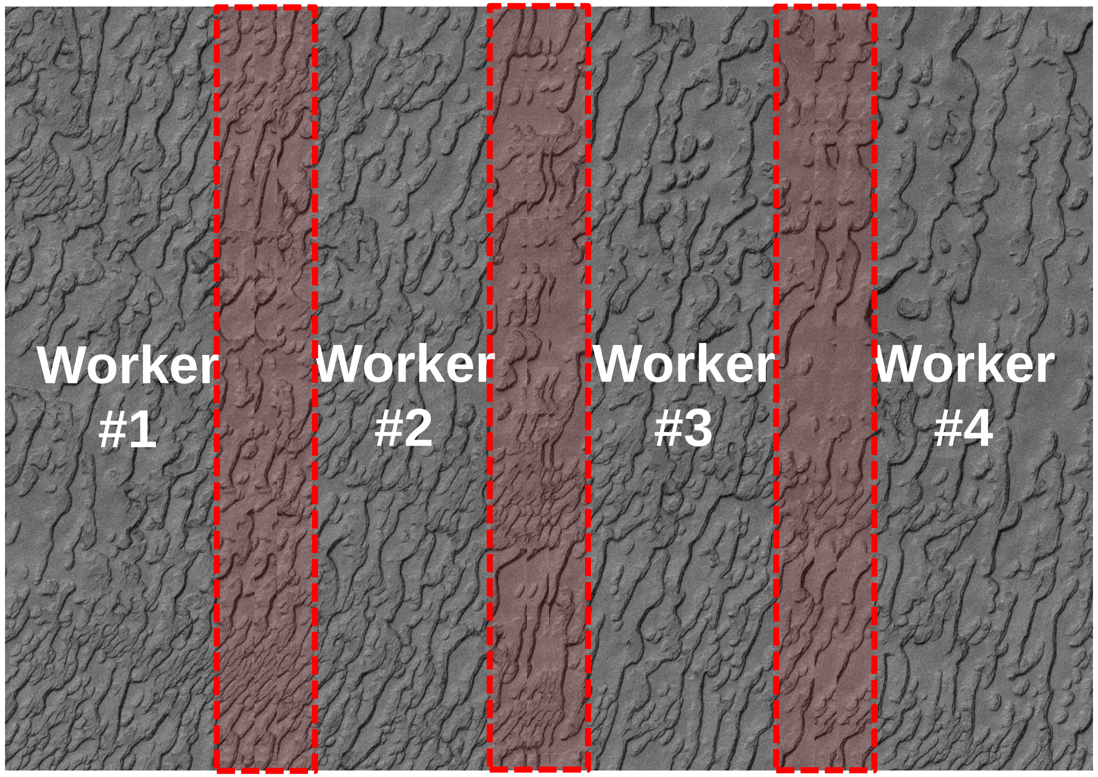
- Divide the image in multiple stripes, each processed by a different worker
- Need for overlay pixels due to the filter's algorithm



Communicating tasks – Denoising a Hi-Res Image

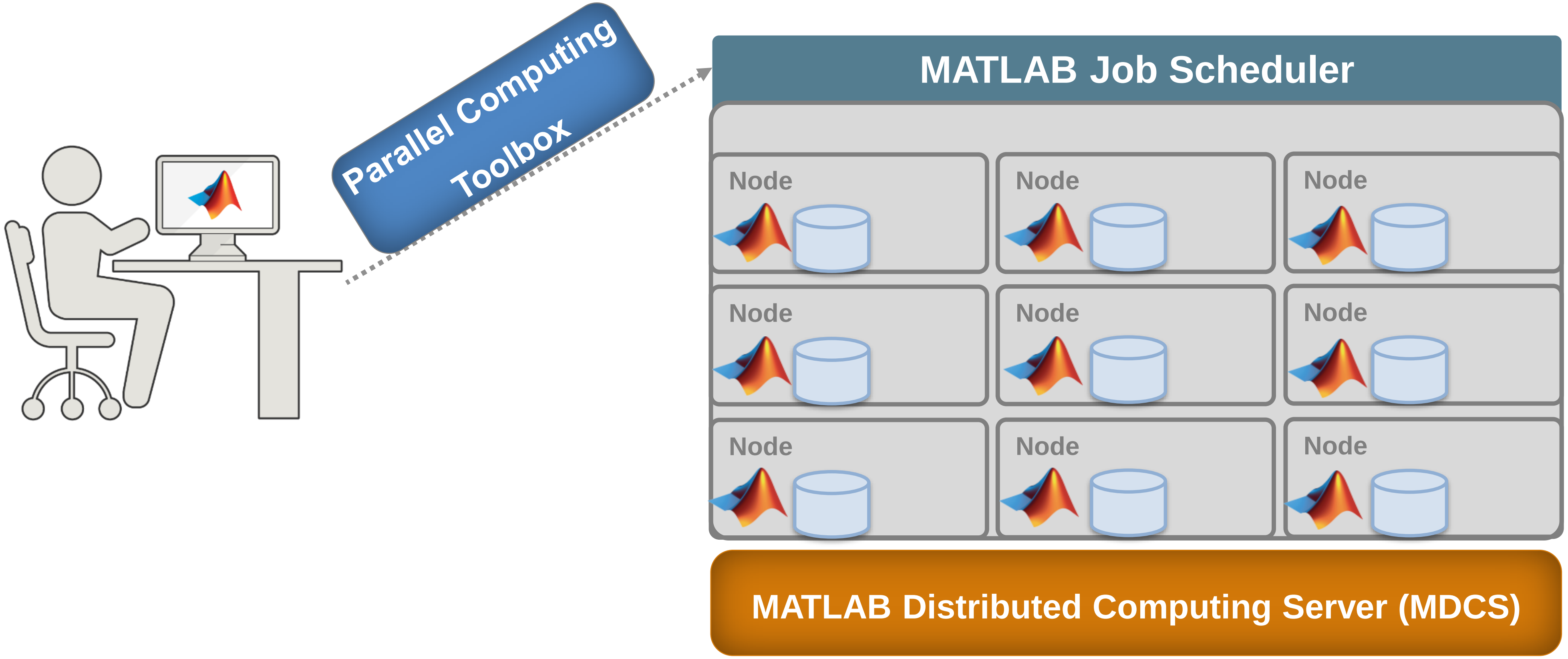
Parallelization technique on multicore CPU(s)

- Divide the image in multiple stripes, each processed by a different worker
- Need for overlay pixels due to the filter's algorithm
- Once processed, reconcatenate the stripes



Running jobs interactively on the cluster

```
cluster = parcluster('MyCluster');
parpool(cluster);
```



Integration with third-party schedulers

- MATLAB Distributed Computing Server includes a MATLAB job scheduler and support for other the following schedulers: SLURM*, LSF, HPC server, PBS and other generic schedulers.

- **Submission Mode with generic schedulers:**

  - **Shared** — When the client machine is able to submit directly to the cluster and there is a shared file system present between the client and the cluster machines.
  - **Remote Submission** — When there is a shared file system present between the client and the cluster machines, but the client machine is not able to submit directly to the cluster (for example, if the scheduler's client utilities are not installed).
  - **Nonshared** — When there is not a shared file system between client and cluster machines.

*** Customizable integration uses the generic scheduler interface.**

What about scalability to an increasing volume of data (Big Data)?

Big Data workflow

1 ACCESS

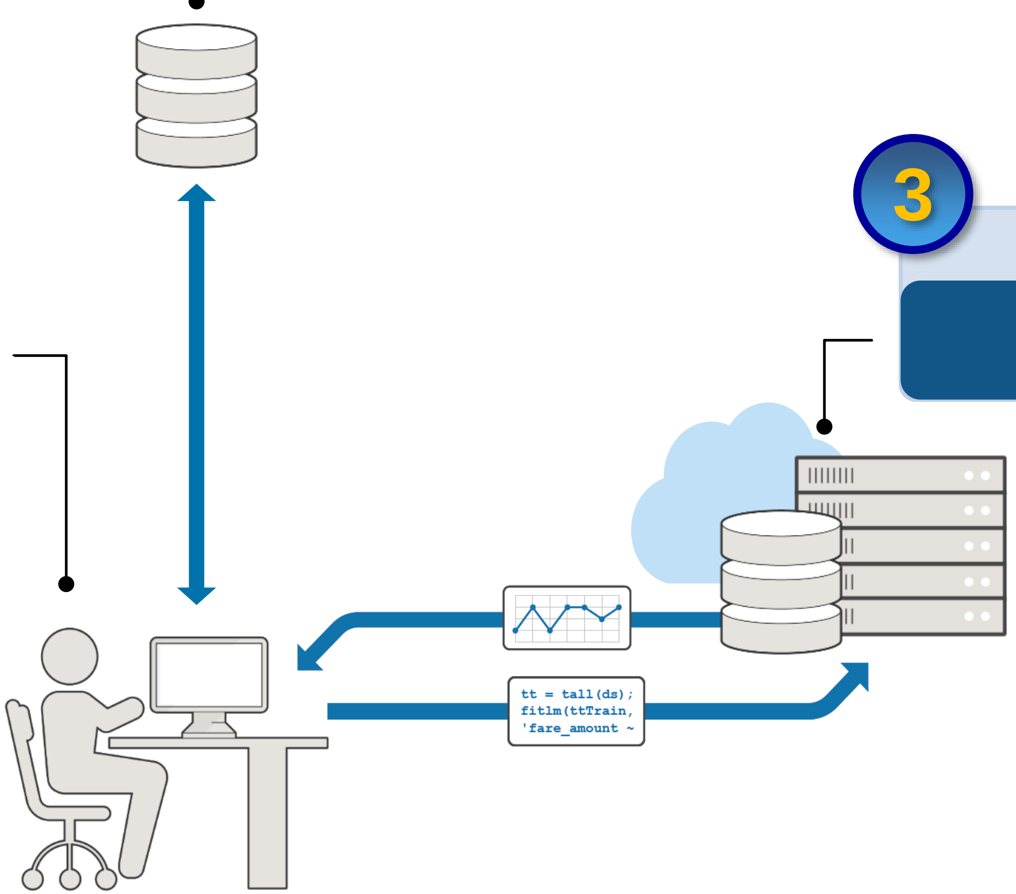
More data and collections of files that do not fit in memory

2 PROCESS AND ANALYZE

Adapt traditional processing tools or learn new tools to work with Big Data

3 SCALE AND DEPLOY

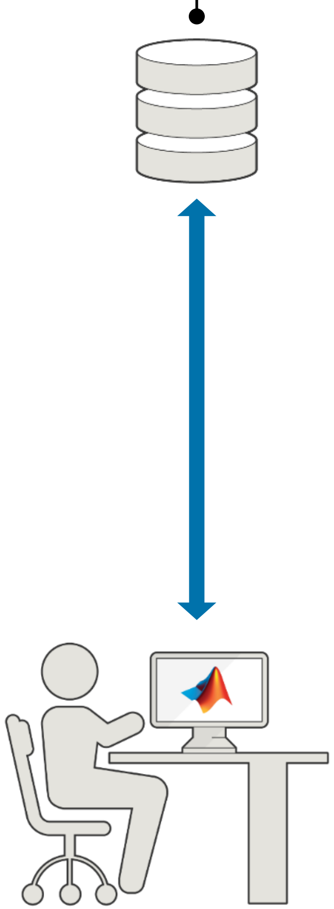
To Big Data systems like Hadoop/Spark



Big Data capabilities in MATLAB

ACCESS

Access data and collections of files that do not fit in memory



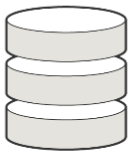
Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

Big Data capabilities in MATLAB

ACCESS

Access data and collections of files that do not fit in memory



Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

- Matrix Math
- Image Processing

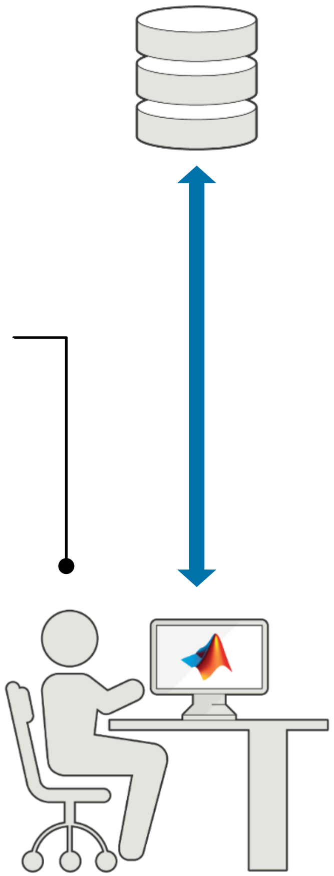
Big Data capabilities in MATLAB

ACCESS

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)



PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

- Matrix Math
- Image Processing

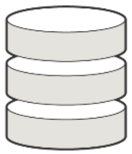
MapReduce

- Statistics

Big Data capabilities in MATLAB

ACCESS

Access data and collections of files that do not fit in memory



Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

- Matrix Math
- Image Processing

MapReduce

- Statistics

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning

Big Data capabilities in MATLAB

ACCESS

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

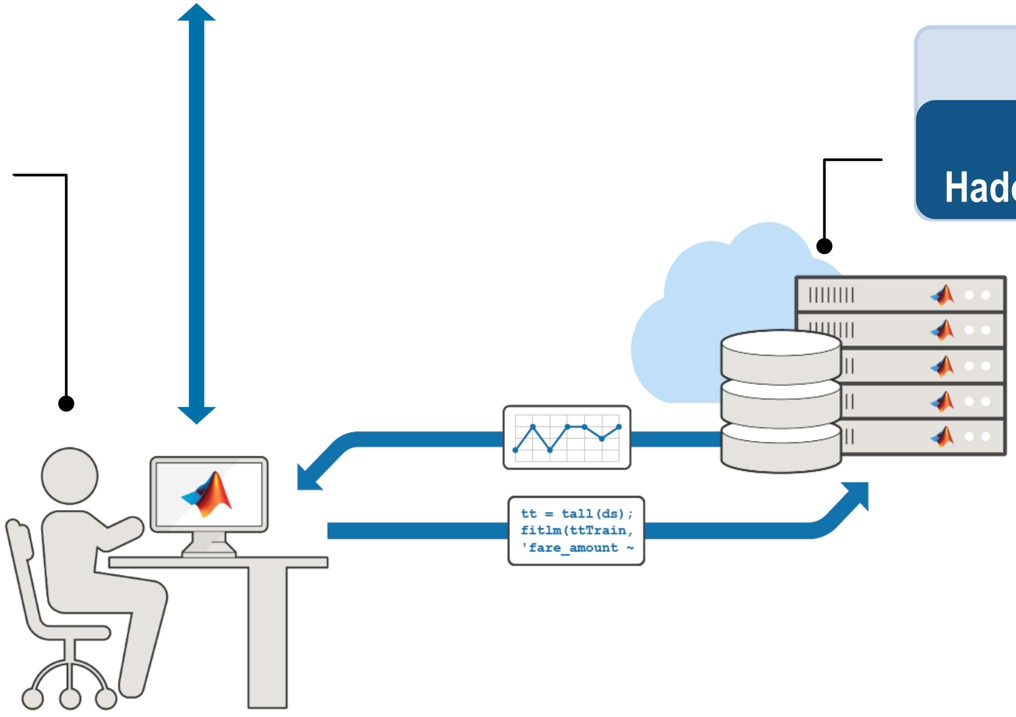
- Matrix Math
- Image Processing

MapReduce

- Statistics

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning



SCALE AND DEPLOY

To compute clusters and Hadoop/Spark for data stored in HDFS

Distributed Arrays

- Matrix Math on compute clusters

Big Data capabilities in MATLAB

ACCESS

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

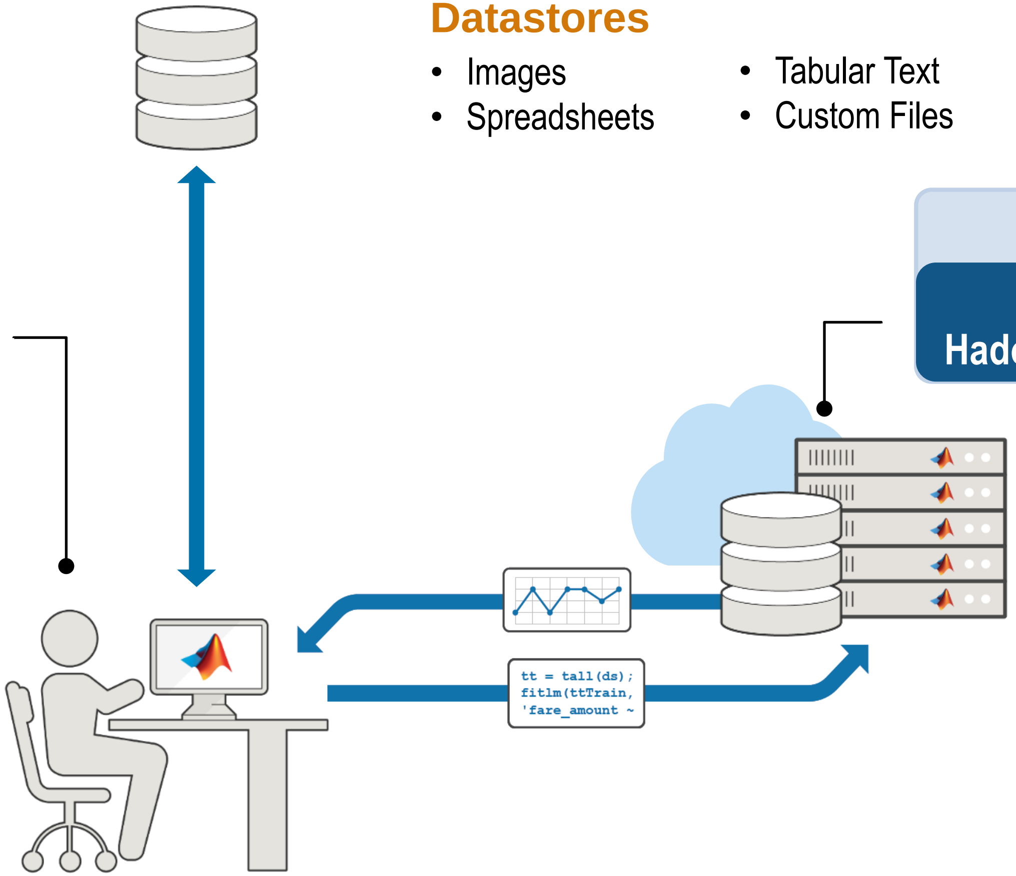
- Matrix Math
- Image Processing

MapReduce

- Statistics

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning



SCALE AND DEPLOY

To compute clusters and Hadoop/Spark for data stored in HDFS

Distributed Arrays

- Matrix Math on compute clusters

MapReduce on cluster

- Compute cluster or Hadoop plateform

Big Data capabilities in MATLAB

ACCESS

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

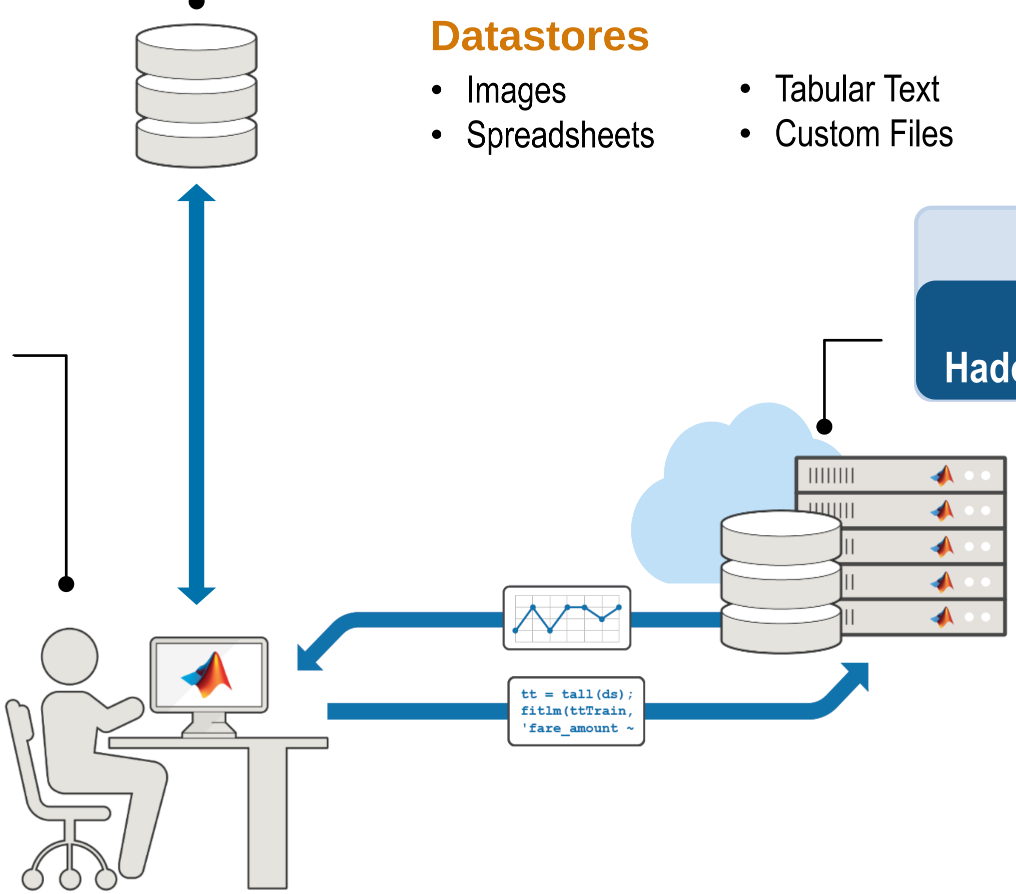
- Matrix Math
- Image Processing

MapReduce

- Statistics

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning



SCALE AND DEPLOY

To compute clusters and Hadoop/Spark for data stored in HDFS

Distributed Arrays

- Matrix Math on compute clusters

MapReduce on cluster

- Compute cluster or Hadoop plateform

Tall Arrays/Spark

- Math, Stats, Machine Learning

Big Data capabilities in MATLAB: Tall Arrays Workflow

ACCESS

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

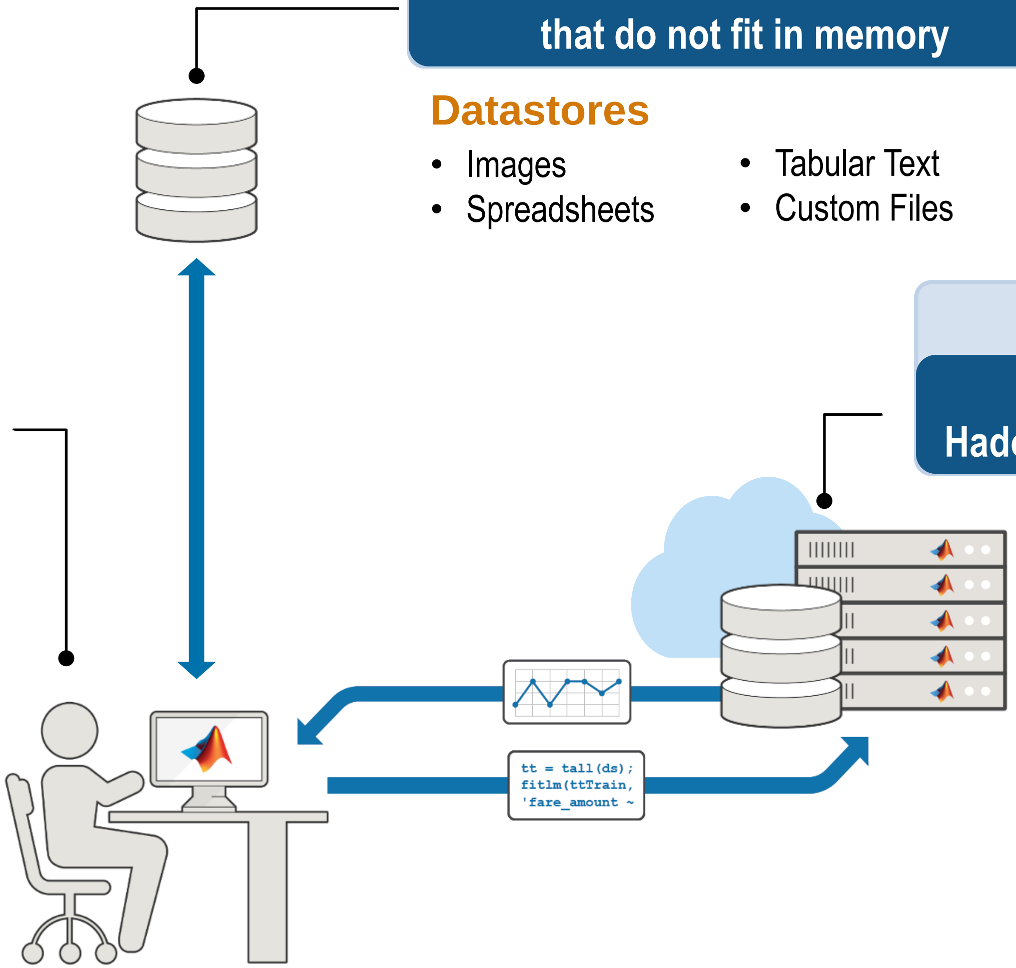
- Matrix Math
- Image Processing

MapReduce

- Statistics

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning



SCALE AND DEPLOY

To compute clusters and Hadoop/Spark for data stored in HDFS

Distributed Arrays

- Matrix Math on compute clusters

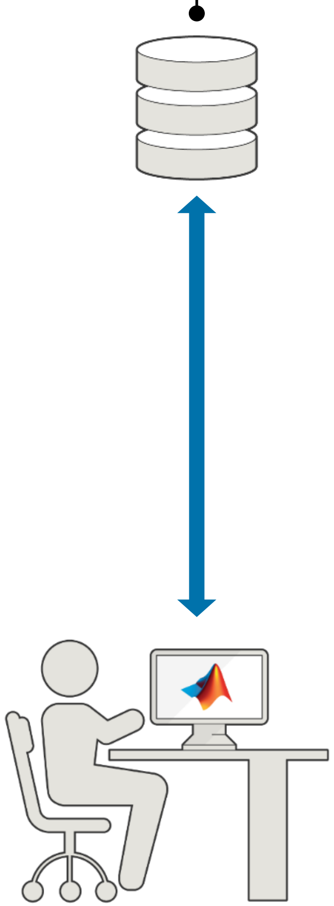
MapReduce on cluster

- Compute cluster or Hadoop plateform

Tall Arrays/Spark

- Math, Stats, Machine Learning

Big Data capabilities in MATLAB



ACCESS

Access data and collections of files that do not fit in memory

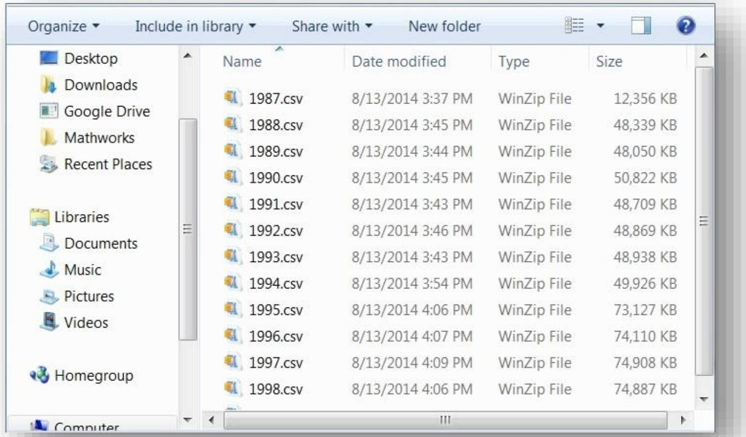
Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

Access Big Data

`datastore`

- Easily specify data set
  - Single text file (or collection of text files)
  - Hadoop HDFS
  - Images
  - Database (SQL, No SQL,)
- Preview data structure and format
- Select data to import using column names
- Incrementally read subsets of the data




```
>> preview(ds)
ans =
      Year      Month      DayofMonth      DayOfWeek
      ____      _____      __________      _________
      1987      10         21              3
      1987      10         26              1
      1987      10         23              5
      1987      10         23              5
```

```
airdata = datastore('*.csv');
airdata.SelectedVariables = {'Distance', 'ArrDelay'};

data = read(airdata);
```

Example: Access Big Data with MATLAB

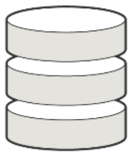
VendorID,	tpep_pickup_datetime,	tpep_dropoff_datetime,	passenger_count,	trip_distance,	pickup_longitude,	picku
2,	2015-01-07 07:40:20,	2015-01-07 08:04:45,	6,	9.12,	-73.9524536132812,	40.78
2,	2015-01-21 22:49:50,	2015-01-21 23:17:11,	6,	5.63,	-74.0083694458008,	40.73
1,	2015-01-05 23:04:30,	2015-01-05 23:15:00,	1,	2.9,	-73.8632125854492,	40.76
1,	2015-01-11 22:20:43,	2015-01-11 22:23:02,	1,	0.8,	-73.9577560424805,	40.76
2,	2015-01-24 00:34:59,	2015-01-24 00:38:39,	1,	0.65,	-73.9916687011719,	40.73
1,	2015-01-25 19:09:57,	2015-01-25 19:18:02,	1,	1.5,	-73.9983825683594,	40.72
1,	2015-01-02 23:24:13,	2015-01-02 23:27:30,	1,	1,	-73.9963912963867,	40.75
2,	2015-01-21 06:46:23,	2015-01-21 06:47:56,	1,	0.63,	-73.9913635253906,	40.77
2,	2015-01-23 19:32:33,	2015-01-23 19:49:56,	3,	2.52,	-73.999382019043,	40.73

- **Objective:** Access data and calculate the average trip duration of New York taxi.
- **Description:**
  - Monthly taxi ride log files
  - The full data set is **big** (~11 GB)
- **Approach:**
  - Preprocess data
  - Calculate average trip duration

Big Data capabilities in MATLAB: Integration with SPARK

ACCESS

Access data and collections of files that do not fit in memory



Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

- Matrix Math
- Image Processing

MapReduce

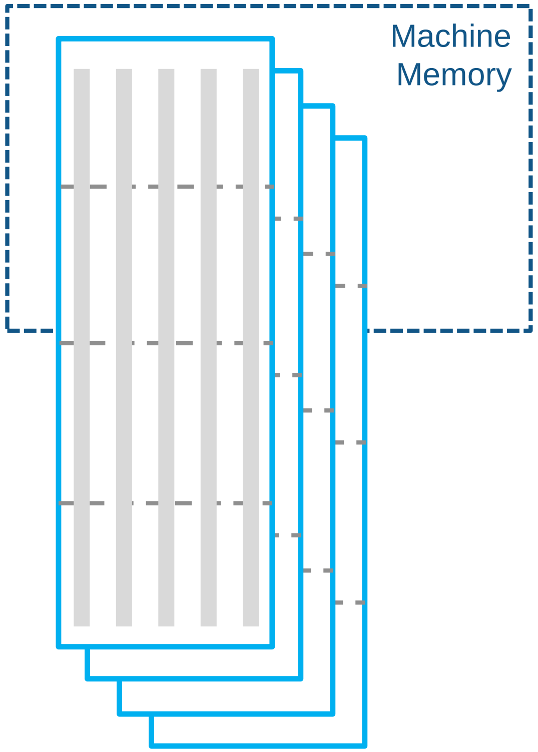
- Statistics

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning

Analyze Big Data using Tall Arrays

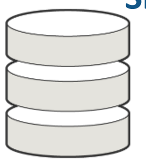
- New data type in MATLAB
- Applicable when:
  - Data is **columnar** – with **many** rows
  - Overall data size is **too big to fit into memory**
  - Operations are mathematical/statistical in nature
- Statistical and machine learning applications
  - Hundreds of functions supported in MATLAB and Statistics and Machine Learning Toolbox



Tall Data

- Prototype your algorithms on desktop to easily scale your big data applications to compute clusters and clusters with Hadoop/Spark

What actually happens with tall arrays in MATLAB?



- Datastore
  - Works as it always has

```
>> fileLoc = 'alltaxiData\*.csv';
>> ds = datastore(fileLoc);
```

- Tall array
  - Pointer to the data; not actually available in memory

```
>> tt = tall(ds);
```

- Functions
  - Operations are queued up; execution is deferred

```
>> tt.mins = minutes(tt.dropoff-tt.pickup);
```

```
>> mn = gather(tt.mins)
```

- Gather step
  - Function that forces execution and gathers results back to the workspace *(result must fit in memory)*
  - Happens with `gather` function, visualizations, and fitting/modeling functions

Example: Tall Arrays

- **Objective:** Create a model to predict the cost of a taxi ride in New York City

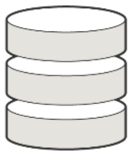
- **Inputs:**
  - Monthly taxi ride log files
  - The local data set is **small** (~2 MB)

- **Approach:**
  - Preprocess and explore data
  - Develop and validate predictive model (linear fit)
    - Work with this subset of data for prototyping

Big Data capabilities in MATLAB: Integration with SPARK

ACCESS

Access data and collections of files that do not fit in memory



Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Parallel computing (CPU and GPU)

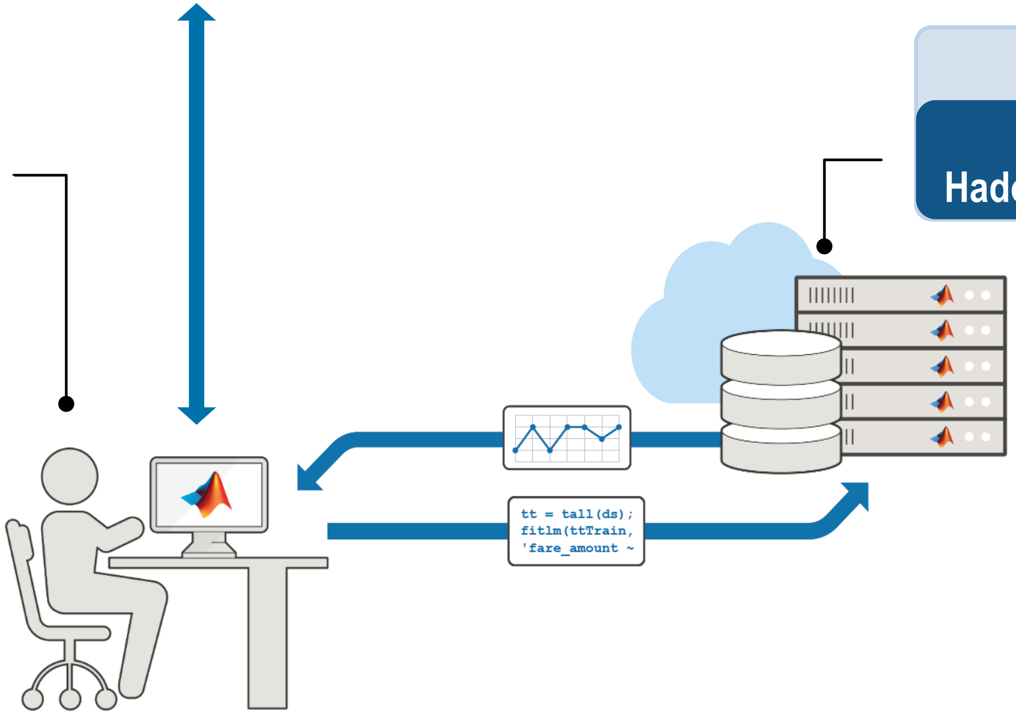
- Matrix Math
- Image Processing

MapReduce

- Statistics

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning



SCALE AND DEPLOY

To compute clusters and Hadoop/Spark for data stored in HDFS

Distributed Arrays

- Matrix Math on compute clusters

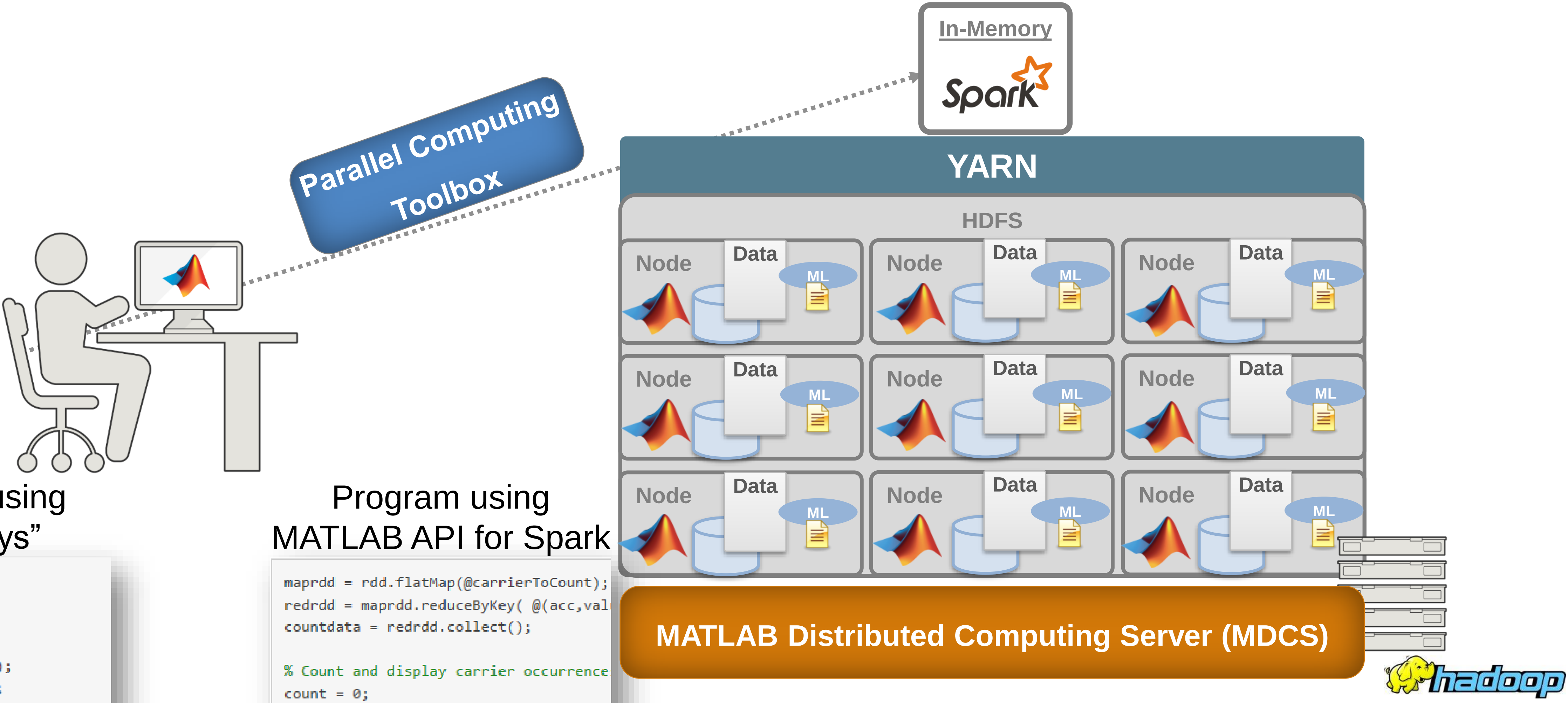
MapReduce on cluster

- Compute cluster or Hadoop plateform

Tall Arrays/Spark

- Math, Stats, Machine Learning

Analyze Big Data on Spark enabled Hadoop cluster



Program using "tall arrays"

```
tt = tall(ds)

a = tt.ArrDelay;

m = mean(a,'omitnan');
s = std(a,'omitnan');
one_sigma_bounds = [m-s m m+s];

sig1 = gather(one_sigma_bounds)
```

Program using MATLAB API for Spark

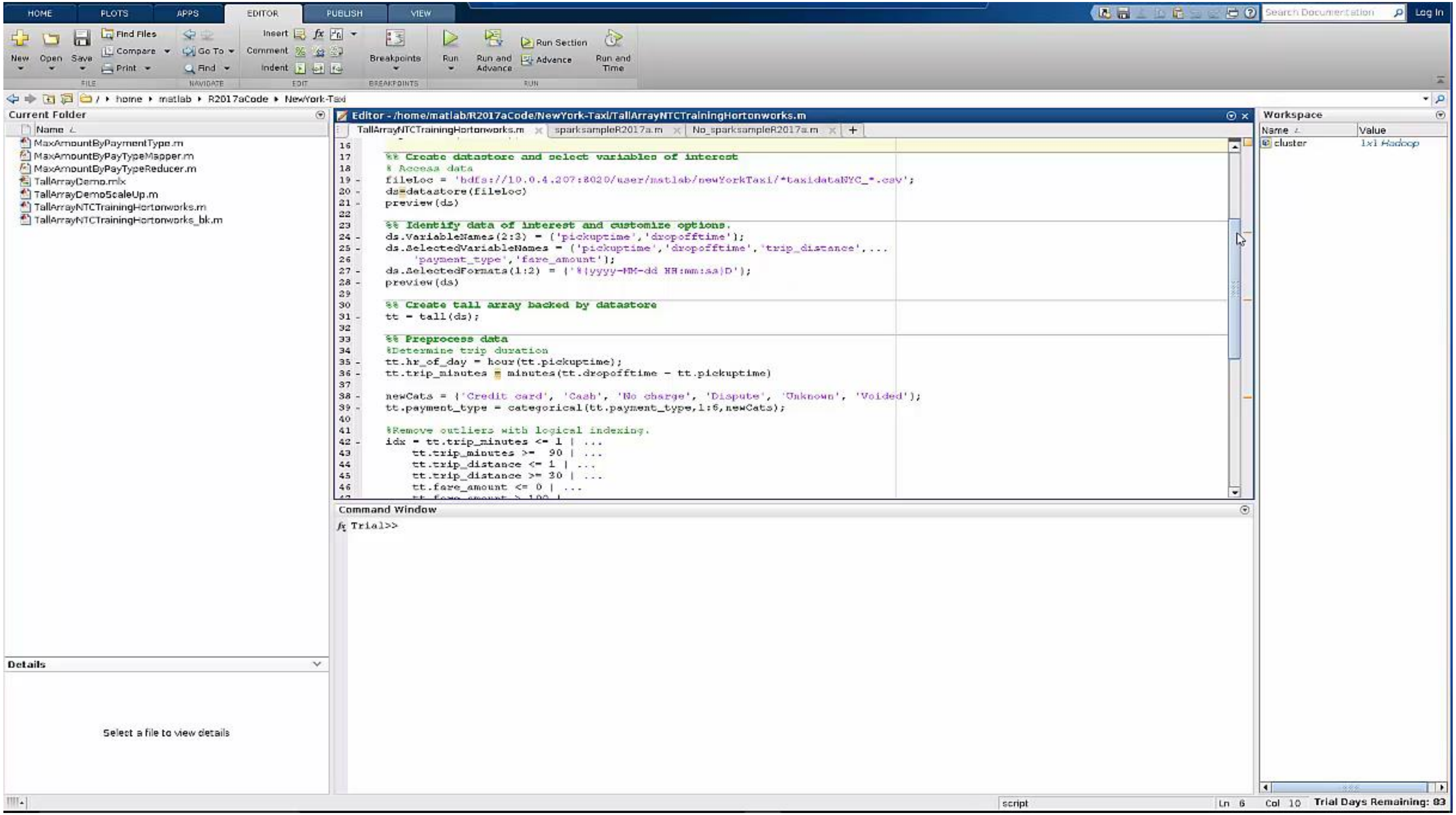
```
maprdd = rdd.flatMap(@carrierToCount);
redrdd = maprdd.reduceByKey( @(acc,val
countdata = redrdd.collect();

% Count and display carrier occurrence
count = 0;
for i=1:numel(countdata)
    count = count + out{i}{2};
```

Example: Working with Big Data on a Spark enabled Hadoop Cluster

- **Objective:** Create a model to predict the cost of a taxi ride in New York City
- **Inputs:**
  - Monthly taxi ride log files
  - The full data set is **big** (~11 GB)
- **Approach:**
  - Preprocess and explore data
  - Develop and validate predictive model (linear fit)
    - Use the algorithm prototype developed with Tall Arrays to scale to the full data set on HDFS

Example: Running on Spark + Hadoop



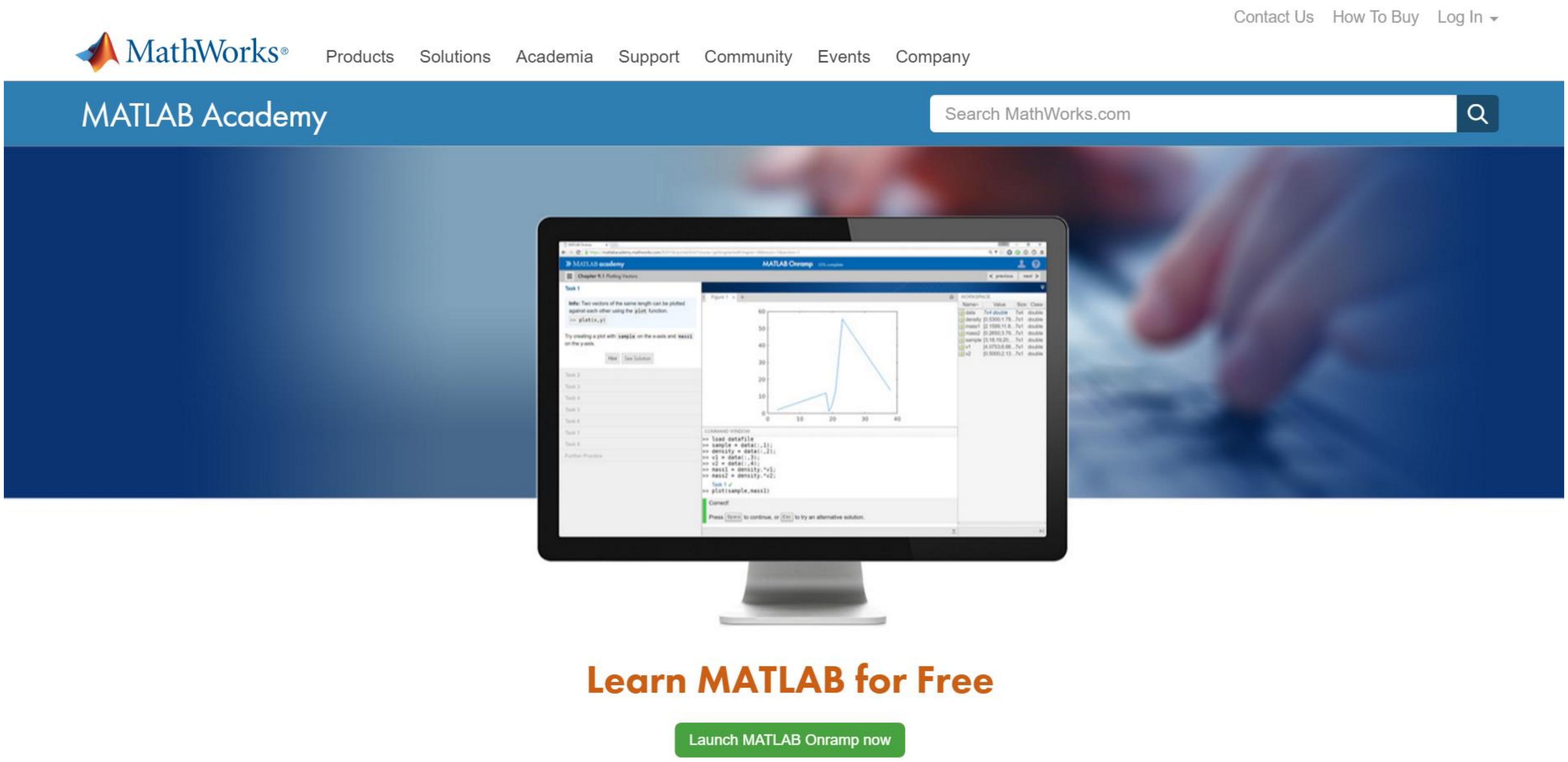
Accès à MATLAB

- Accès au réseau du laboratoire
  - Nombre de jetons illimité
  - Accès à toutes les toolboxes de la licence campus

- Accès sur PC personnel

Uniquement aux étudiants et au personnel de l'Université d'Orléans
prenom.nom@univ-orleans.fr
prenom.nom@etu.univ-orleans.fr

Self-paced MATLAB Tutorials

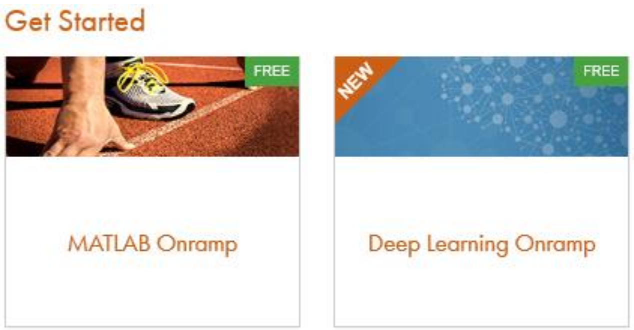


http://matlabacademy.mathworks.com

Search: MATLAB Academy

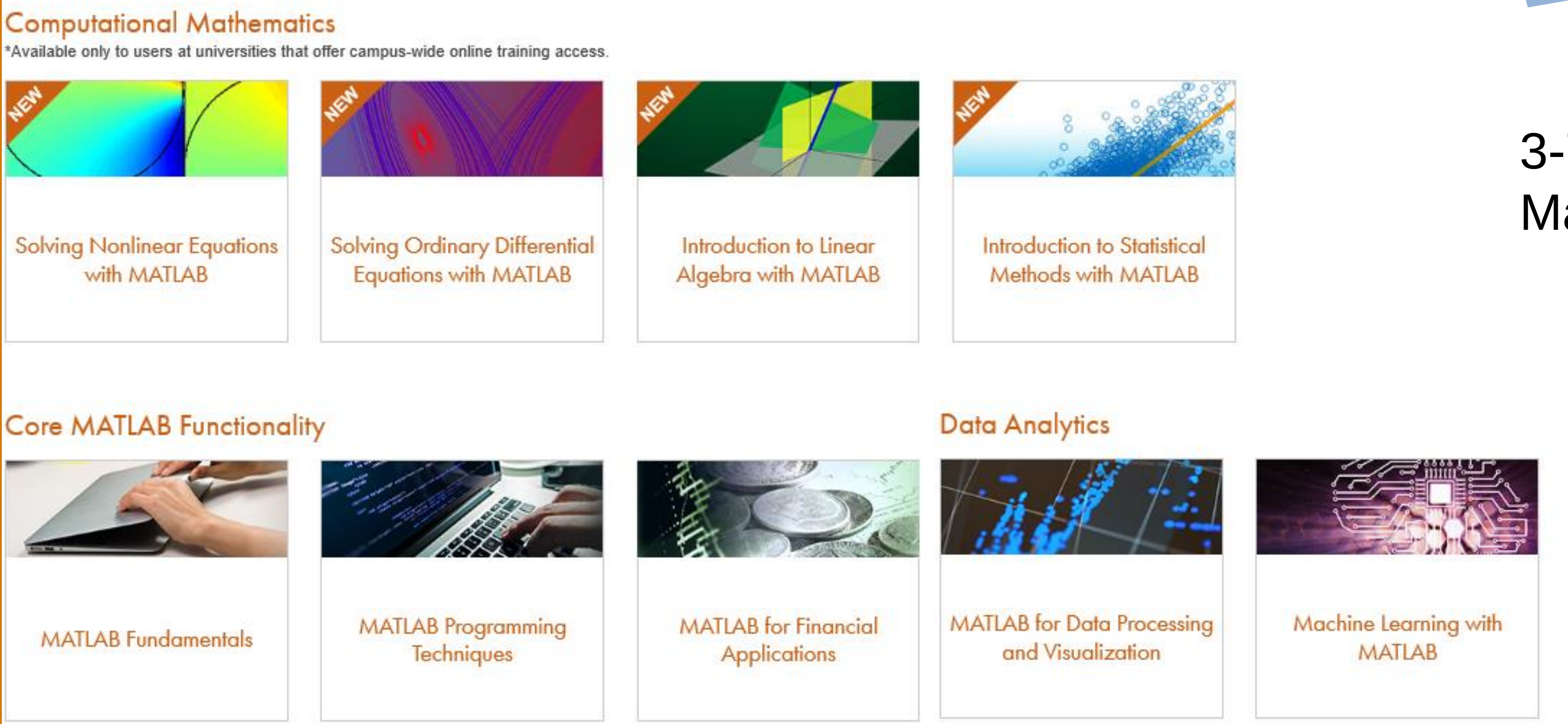
Formations en ligne

Pour tous



2- 4 hours content

Uniquement aux étudiants et au personnel de l'Université d'Orléans
prenom.nom@univ-orleans.fr
prenom.nom@etu.univ-orleans.fr



3-7 hours of solving Computational Mathematics with MATLAB

>80 hours of learning how to improve MATLAB skills

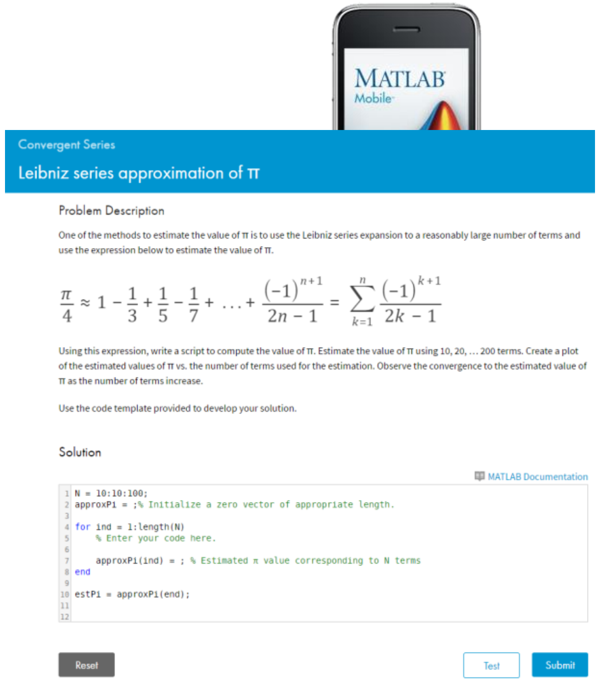
Objectifs: Présentation des outils MathWorks pour la pédagogie mis en place pour

Enseignant

- Economiser du temps pour les enseignants

Etudiant

- Motiver et accélérer l'apprentissage des étudiants




```
N = 10:10:100;
approxPi = ;% Initialize a zero vector of appropriate length.

for ind = 1:length(N)
    % Enter your code here.

    approxPi(ind) = ; % Estimated π value corresponding to N terms
end

estPi = approxPi(end);
```

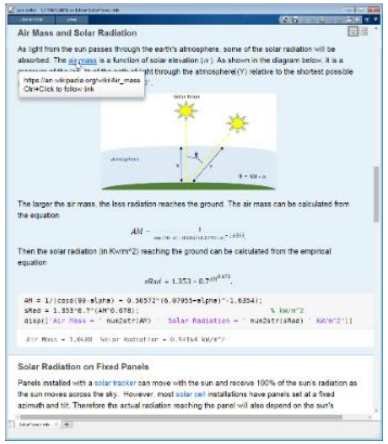
Formations en ligne

MATLAB mobile

Live Script editor

Cody Coursework

Connectivité Hardware

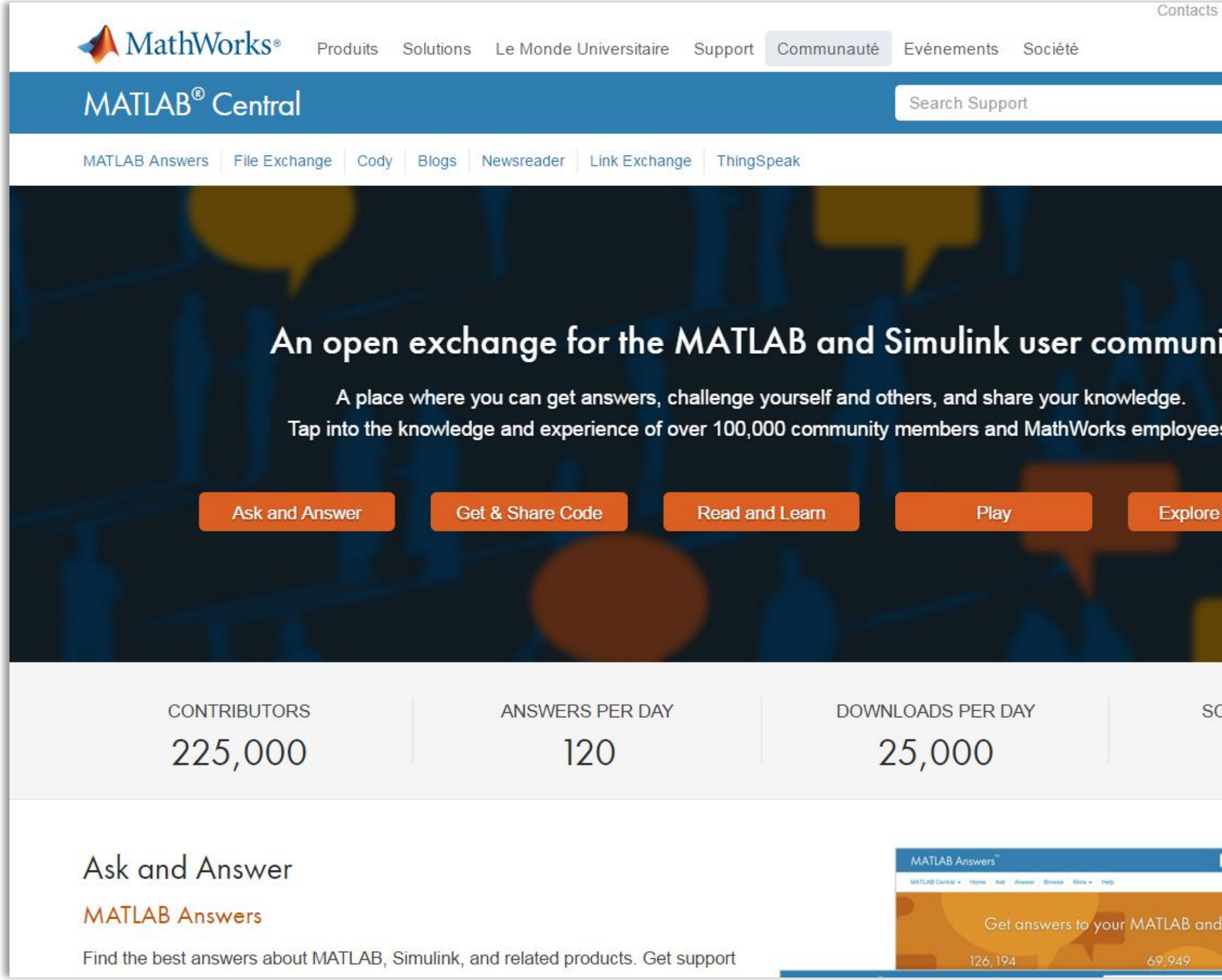


Summary

- Easily develop parallel MATLAB applications without being a parallel programming expert
- Speed up the execution of your MATLAB applications using additional hardware
- Develop parallel applications on your desktop and easily scale to a cluster when needed
- Scale-up your MATLAB applications developed locally to Big Data systems

MATLAB Central

- Community for MATLAB and Simulink users
- File Exchange
  - Access free files including functions, apps, examples, and models
- MATLAB Answers
  - Ask programming questions or search community-answered questions
- Newsgroup
  - Participate in technical discussions
- Blogs
  - Read commentary from engineers who design, build, and support MathWorks products



Merci

Questions?